

PR #29715 完整报告

sgl-project/sglang

[CI] Migrate JIT tests missed by #29066 to runner_config registration

合并时间: 2026-06-30 14:21

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29715>

执行摘要

- 一句话: 补齐 JIT 测试 CI 注册迁移并加固验证规则
- 推荐动作: 值得精读 `check_registered_tests.py` 的改动, 了解如何设计防御性 CI 审核工具以防止回归。其他测试文件的迁移模式可作为后续 CI 注册的标准范例。

功能与动机

修复了 #29066 迁移过程中因 squash merge 不重新扫描 main 导致的测试静默丢弃问题。PR body 指出: 'During that ~4-day review window, 9 other JIT-kernel PRs merged to main, each registering with the old suite="base-b-kernel-..." convention. ... their effective_suite stayed in the old no-test- shape ... and no longer matches any suite the workflow invokes — so they were silently dropped from PR CI.'

实现拆解

1. 迁移漏掉的测试注册: 修改 `test/registered/jit/` 和 `test/registered/jit/benchmark/` 下的 15 个文件, 将 `register_cuda_ci` 调用从 `suite=` 形式改为 `stage=/runner_config=` 形式; 同时将三个 B200 注册从已移除的 1-gpu-b200 重定向到标准化的 4-gpu-b200 runner。
2. 加固验证规则: 在 `scripts/ci/check_registered_tests.py` 中增加 `_LEGACY_CUDA_PREFIXES` 元组, 只允许 `nightly/stress/weekly` 前缀的套件继续使用旧 `suite=` 形式; 其他 CUDA `suite=` 若为现代 `{stage}-test-{runner_config}` 形状则提示改用 `stage=/runner_config=`, 若为旧 `{stage}-{runner_config}` 形状则直接报错。确保 pre-commit 和 CI 能捕获任何遗漏的旧格式。
3. 重新归类分布式通信测试: 将 `test/registered/dcp/test_reduce_scatter_along_dim.py` 从 `base-b-kernel-unit-8-gpu-h200` 迁移到 `extra-b/8-gpu-h200`, 使其匹配正确的工作流 (`pr-test-extra.yml`), 因为该测试不测试内核, 而是测试 `GroupCoordinator.reduce_scatter_along_dim`。

关键文件:

- `scripts/ci/check_registered_tests.py` (模块 CI 审核; 类别 infra; 类型 infrastructure; 符号 `_LEGACY_CUDA_PREFIXES`, `main`): 核心基础设施变更, 加强 CI 注册验证规则, 防止旧格式再次漏过。
- `test/registered/jit/test_dsv3_router_gemm.py` (模块 内核测试; 类别 test; 类型 test-coverage): 作为 JIT 单元测试迁移的代表, 展示从 `suite=` 到

stage=/runner_config= 的转换。

- test/registered/dcp/test_reduce_scatter_along_dim.py (模块 通信测试; 类别 test; 类型 test-coverage) : 被重新归类的测试, 从错误的 kernel 套件移动到正确的 extra-b 工作流, 避免测试被静默丢弃。

关键符号: check_registered_tests.main

关键源码片段

scripts/ci/check_registered_tests.py

核心基础设施变更, 加强 CI 注册验证规则, 防止旧格式再次漏过。

```
# scripts/ci/check_registered_tests.py 核心验证逻辑
import re

# 现代 suite 名称格式: {stage}-test-{runner_config}
_MODERN_SHAPE = re.compile(r"^(.+)-test-(.+)")

# 允许继续使用遗留 suite= 的前缀白名单 (夜间 / 压力 / 每周套件)
_LEGACY_CUDA_PREFIXES = ("nightly", "stress", "weekly")

def main() -> int:
    # 解析所有测试文件, 检查注册调用
    missing = []
    legacy_shape = [] # 已使用现代形状但通过旧参数传入 (可自动转换)
    non_dispatchable = [] # 旧形状不可调度, 必须报错

    for f in files:
        registries, _ = ci_register.ut_parse_one_file(f)
        for r in registries:
            if r.suite is None:
                continue
            # 如果已经使用 stage= / runner_config= 现代形式, 跳过
            if r.stage is not None:
                continue
            # 属于 nightly/stress/weekly 系列, 允许保留旧形式
            if r.suite.split("-", 1)[0] in _LEGACY_CUDA_PREFIXES:
                continue
            m = _MODERN_SHAPE.match(r.suite)
            if m:
                # 形状是新式但用了旧参数, 提示作者改用 stage=/runner_config=
                legacy_shape.append((f, r.suite, m.group(1), m.group(2)))
            else:
                # 形状完全不可调度, 直接报错
                non_dispatchable.append((f, r.suite))

    # 报告 legacy_shape 和 non_dispatchable, 返回非零退出码
    if legacy_shape:
        print("ERROR: 以下文件使用了现代 suite 形状但通过旧参数传入, 缺少 /rerun-test 支持: ")
```

```
for f, suite, stage, rc in legacy_shape:
    print(f" {f}: suite={suite!r} \u2192 \u6539\u4e3a stage={stage!r}, runner_config={rc!r}")
if non_dispatchable:
    print("ERROR: 以下文件使用了不可调度的旧 CUDA suite: ")
    for f, suite in non_dispatchable:
        print(f" {f}: suite={suite!r} \u65e0\u6cd5\u5339\u914d\u4efb\u4f55 PR
            \u5de5\u4f5c\u6d41")
    return int(bool(legacy_shape or non_dispatchable))
```

评论区精华

本 PR 无 review 评论。作者通过 PR body 详细说明了问题根因和修复方案，并通过 /rerun-test 触发 CI 验证了所有受影响测试的正确调度。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。主要风险：1) 可能还有其他文件使用了旧格式但未被迁移——新验证规则会捕获；2) B200 runner 从 1-gpu-b200 改为 4-gpu-b200 可能与实际 runner 配置不一致，但作者已确认 4-gpu-b200 是标准名称；3) 重新归类可能导致测试在错误工作流中运行——作者已验证 effective_suite 正确。CI 自动运行确认全部通过。
- 影响：对开发者：PR CI 不再静默丢弃 JIT 测试，所有注册的测试都会按预期运行；新的 pre-commit 检查提前捕获违规格式。对用户：无直接影响。对 CI 系统：测试执行成本略有增加（恢复运行原来被丢弃的测试），但均为已存在测试。
- 风险标记：CI 测试静默丢弃，B200 runner 名称变更，基础设施配置错误

关联脉络

- PR #29066 [CI] Migrate JIT tests to runner config registration: 本 PR 是 #29066 的补丁，迁移了其合并窗口期间漏掉的测试注册。
- PR #14194 [Dist] Add reduce_scatter_along_dim test: 重新归类的 reduce_scatter_along_dim 测试最初来自该 PR。